# PATENT ABSTRACTS OF JAPAN

(11)Publication number : 06-012323
(43)Date of publication of application : 21.01.1994

(51)Int.Cl. G06F 12/08
G06F 12/08

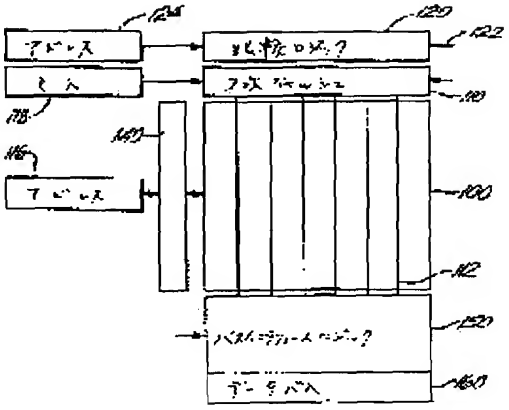
(21)Application number : 05-063469
(22)Date of filing : 26.02.1993

(71)Applicant : HEWLETT PACKARD CO <HP>
(72)Inventor : RAJENDRA KUMAR
EMERSON PAUL G

(30)Priority
Priority number : 92 842907 Priority date : 27.02.1992 Priority country : US

## (54) CACHE MEMORY SYSTEM

(57)Abstract:
PURPOSE: To escape slashing, and to improve a cache hit ratio while making the best use of the original merits of a direct mapping type primary cache by providing a cache memory system in which an integrated secondary cache can be executed with the direct mapping type primary cache.
CONSTITUTION: A complete associative secondary cache 110 is preferably provided close to a direct mapping type primary cache 100. A bus interface logic 150 of the both caches is commonly used so that data can be transmitted to a microprocessor without being transmitted to the primary cache 100 even when the primary cache 100 is miss, and the secondary cache 110 is hit, and through-put can be improved. Also, slashing can be escaped, and a fault at the time of using the conventional set associative cache can be prevented.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平6－12323

(43)公開日　平成６年(1994)１月21日

| (51)Int.Cl.[5] | 識別記号 | 庁内整理番号 | ＦＩ | 技術表示箇所 |
|---|---|---|---|---|
| Ｇ０６Ｆ　12/08 | | Ｆ　7608－5Ｂ | | |
| | ３１０ | 7608－5Ｂ | | |

審査請求　未請求　請求項の数３(全 11 頁)

(21)出願番号　特願平5－63469

(22)出願日　平成５年(1993)２月26日

(31)優先権主張番号　８４２，９０７

(32)優先日　1992年２月27日

(33)優先権主張国　米国（ＵＳ）

(71)出願人　590000400
ヒューレット・パッカード・カンパニー
アメリカ合衆国カリフォルニア州パロアルト　ハノーバー・ストリート　3000

(72)発明者　ラジェンドラ・クマー
アメリカ合衆国カリフォルニア州サニーベイル、ロンドンデリイ・ドライブ　756

(72)発明者　ポール・ジー・エマーソン
アメリカ合衆国カリフォルニア州サンノゼ、アルマーデン・バレイ・ドライブ　1481

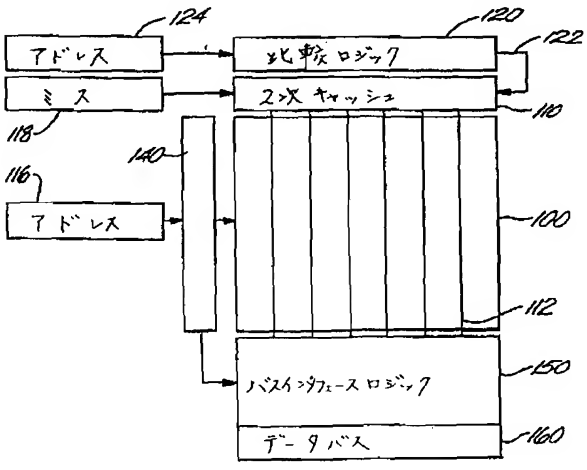
(74)代理人　弁理士　長谷川　次男

(54)【発明の名称】　キャッシュメモリシステム

(57)【要約】

【目的】直接マッピング型１次キャッシュと共に集積された２次キャッシュを伴うキャッシュメモリシステムを提供することで、直接マッピング型キャッシュ本来の長所を生かしながら、スラッシングを回避し、キャッシュヒット率を改善する。

【構成】直接マッピング型１次キャッシュ100 に好ましくは完全連想型２次キャッシュ110 を近接して設ける。この構成によると両キャッシュのバスインターフェースロジック150 が共通となるため、１次キャッシュがミスで２次キャッシュがヒットのときもデータを１次キャッシュに送らずにマイクロプロセッサに送ることができスループットの改善ができる。スラッシングも回避でき、従来セット連想型キャッシュを使っていた際の欠点も防ぐことができる。



1

【特許請求の範囲】

【請求項１】以下の(a)及び(b)を有するキャッシュメモリシステム：

(a)1次キャッシュのヒット／ミスの比較手段と前記比較結果を出力する手段を有する直接マッピング型1次キャッシュ。

(b)前記1次キャッシュと共通の出力データビット線を有する2次キャッシュ。

【請求項２】請求項１記載のキャッシュメモリシステムの2次キャッシュにおいて、前記1次キャッシュのヒット／ミスの比較と同時に前記2次キャッシュのヒットの比較をする手段、前記2次キャッシュの比較結果を出力する手段および読み出しにおいて前記１次キャッシュと同時に読み出す手段を有することを特徴とするキャッシュメモリシステム。

【請求項３】以下の(a)及び(b)を設けたキャッシュメモリシステム：

(a)1次キャッシュは1次記憶と、アドレス入力と前記1次記憶を比較する手段を有する直接マッピング型1次キャッシュ：前記比較手段の結果が一致であれば1次キャッシュヒット信号を発生し：また、前記比較手段の結果が不一致であれば1次キャッシュミス信号を発生する。

(b)前記1次キャッシュに接続され、読み出しにおいて前記1次キャッシュと同時に読めるように構成された完全連想型2次キャッシュ：前記2次キャッシュは2次記憶と、アドレス入力と前記2次記憶を前記1次比較手段と同時に比較する：前記比較手段の結果が一致であれば2次キャッシュヒット信号を発生する。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、広義には電子キャッシュメモリ回路のアーキテクチャ及びその記憶内容の処理方法に関するものである。より詳しくは、本発明は、1次キャッシュ(primary cache)に近接させて2次キャッシュ(secondary cache)を設け、1次キャッシュからの読出しミスが起こると同時にこの2次キャッシュにアクセスするようにした超大規模集積(VLSI)回路用のキャッシュメモリアーキテクチャ及び命令キャッシュアーキテクチャに関する。

【０００２】

【従来の技術】一般に、ランダムアクセスメモリ(RAM)集積回路(IC)は、高度なマイクロプロセッサより動作速度がはるかに遅い。マイクロプロセッサの設計者等は、RAMのアクセスタイムが遅いと、プロセッサのスループット及びプログラムの実行速度の増大に対して大きな障害になるという認識を持つに至った。例えば、最新の縮小命令セットコンピュータ(RISC)型マイクロプロセッサでは、理論上1秒間に200万命令を実行するよう20MHz以上のクロック速度が使用されることがある。しかしながら、典型的なRAMのサイクルタイム（応答時間）は150ナノ秒である。従って、マイクロプロセッサがRAMから新しいデータ値または命令を取り出すことが必要な場合、マイクロプロセッサがRAMの応答を待つ間に多くのマイクロプロセッササイクルが浪費されることが起こり得る。

2

【０００３】この問題を解決するため、マイクロコンピュータやマイクロプロセッサでは、キャッシュメモリを用いてメモリアクセスタイムを改善することが行われる。キャッシュメモリは、高速なキャッシュメモリが低速なRAMのアクティブ部分のコピーをそっくり記憶するという点において仮想メモリに類似している。通常、キャッシュメモリは、アクセスタイムが主メモリより4～20倍速くなるようマイクロプロセッサチップ上に配置される。

【０００４】

【発明が解決しようとする課題およびそのための手段】キャッシュメモリの動作は次のようになっている。マイクロプロセッサによってメモリ要求が出されると、その要求はキャッシュメモリに提示され、キャッシュが応答することができないと、その要求は主メモリに提示される。マイクロプロセッサが、キャッシュ中になく、主メモリ中にある項目にアクセスしようとすると、「キャッシュミス」が起こる。キャッシュミスに応答して主メモリからキャッシュに必要なデータをロードすることによって、キャッシュは更新される。次に、それらのデータはキャッシュからマイクロプロセッサへ供給される。所望のデータあるいはデータが入っていないキャッシュ中のラインは「犠牲」ラインと呼ばれる。

【０００５】キャッシュミス時にキャッシュの状態を更新するために用いることのできる時間は非常に短い。そのため、キャッシュは、所要時間内にキャッシュミスを自動的に処理することができるハードウェアによって制御される。不都合なことに、従来技術の回路においては、キャッシュミスは、全て主メモリにアクセスすることによってキャッシュを更新することが必要であり、そのためにマイクロプロセッサのスループットが著しく低くなっていた。従って、本発明の1つの目的は、キャッシュミス後に行われる主メモリアクセス数を少くすることによってスループットを改善することにある。

【０００６】従来、キャッシュは、大別して直接マッピング型、セット連想型及び完全連想型の3種類の構造のものがある。これらの3種類のキャッシュの詳細は、次の従来技術に関する参考文献に記載されている：De Blasi著"Computer Architecture", ISBN 0-201-41603-4(Addison-Wesley、1990年)，273～291ページ；Stone著"High Performance Computer Architecture", ISBN 0-201-51377-3(Addison-Wesley,第2版，1990年)，29～39ページ；Tabak著"Advanced Microprocessors), ISBN 0-07-062807-6(McGraw-Hill，1991年)，244～248ページ。これらの文献は当業者には周知である。

【０００７】これらの3種類のどのキャッシュにおいても、入力アドレスは比較ロジックに供給される。通常は、タグビットと呼ばれるアドレスの部分集合が入力アドレスから抽出され、各キャッシュエントリのタグビットと比較される。これらのタグビットが一致（マッチ）すれば、対応するデータがキャッシュから抽出される。直接マッピング型キャッシュの全体的構成及び処理の仕組を図1に示す。この直接マッピング型キャッシュはキャッシュメモリ10を含み、キャッシュメモリ10は複数のタグ12及びデータ要素14を有するテーブルとして実現することができる。これらのタグ及びデータはペアとしてアクセスされる。入力アドレス20は、マイクロプロセッサからアドレスデコード回路30へ供給され、ここで入力アドレス20よりタグビットを分離する。タグビットは、第1の入力42として比較器40に供給される。また、比較器40は、入力アドレスの下位ビットにより指示されるキャッシュメモリ10の記憶場所からのタグビットを有する第2の入力44も受け取る。このように、下位の入力アドレスビットは、キャッシュメモリ中のタグを一つだけ指示する。比較器40においてこれらのタグビットが互いに一致すると、比較器は、そのヒット出力60を活性化、つまりアサート(assert)し、データ選択回路70にキャッシュメモリからデータ要素14を読み出させる。タグとデータ要素はペアで配置されているから、データ選択回路は一致したタグに対応するデータ要素を受け取る。このように選択されたデータは、キャッシュメモリからマイクロプロセッサへ出力80として供給され、さらに処理される。

【０００８】第1の入力42と入力アドレスの下位ビットによって指示されたキャッシュメモリ中の記憶場所のタグビットが一致しなかった場合は、比較器40はそのミス出力50をアサートする。すると、ミス処理（ブロック55によって表される）がトリガされるが、このミス処理は、大方の従来技術のデバイスの場合、主メモリへのアクセスを必要とする。

【０００９】一般に、直接マッピング型キャッシュは、アクセスは最も速いが、所要時間のほとんどをタグビットの比較に費やす。完全連想型キャッシュは、タグビットの比較は速いが、電力消費が大きく、かつより複雑な回路を必要とする。

【００１０】従来技術においては、キャッシュは「スラッシング」を起こし易かった。スラッシングは、マイクロプロセッサが所望のデータ項目を探索して、それを見付けることができず、キャッシュをそのデータ項目によって更新し、後でその更新したラインを異なる項目に置換するという動作を繰り返す場合に起こる。これにより、同じデータ項目に関してキャッシュミスを繰り返すサイクル動作が引き起こされる。

【００１１】スラッシングを回避するには、通常大きなセット連想型キャッシュが用いられる。しかしながら、この種のキャッシュにはいくつか欠点がある。まず、キャッシュに対する読出しと書込みが、キャッシュタグの比較が皆終わった後に実行される別のマシンサイクルで行われ、キャッシュアクセスタイムが大きくなる。また、これによってマシンサイクルタイムが長くなるか、または遅延サイクルが必要となることにより、マイクロプロセッサ性能も低下する。第2には、例えばn-ウェイのセット連想型キャッシュの場合、n倍多くのワードを読み出した後で、タグ比較に基づき所与のワードの選択を行わなければならないため、キャッシュ中の電力消費が著しく大きくなる。第3に、キャッシュの動作は、複数のセンスアンプが同時にトリガされる際に発生する回路のスイッチングノイズの影響を非常に受け易い。この傾向は、ワードサイズが32ビット以上の機械、及び複数のキャッシュポートを介して複数のワードを同時に読み出すことができるマルチポートキャッシュを有する機械において特に顕著である。

【００１２】従って、本発明の1つの目的は、セット連想型の構成を用いることなくスラッシングを回避することのできるキャッシュを提供することにある。

【００１３】かつてある研究者によって、直接マッピング型の1次キャッシュのバックアップとして犠牲キャッシュまたはミスキャッシュを使用するという考えが提案された。スラッシングを回避し、ヒット率を改善するための2次キャッシュ（「犠牲／ミスキャッシュ」と呼ばれる）のアーキテクチャ上の定義は、コンピュータアーキテクチャに関する第17回年次国際シンポジウムの会報所載のN. Jouppi著"Improving Direct-Mapped Cache Performance by the Addition of a Small Fully-Associative Cache and Prefetch Buffers"(IEEE Computer Society Press, 1990年5月)の364～373ページに記載されている。しかしながら、Jouppiは、VLSIにおける2次キャッシュの具体的使用については何ら教示しておらず、またビット線、センスアンプ及びバスインタフェース・ロジックを共用する形での1次キャッシュと2次キャッシュの一体化についても何ら教示していない。

【００１４】従来、キャッシュメモリは、データ用だけではなく命令キャッシュとしても用いられて来た。従来技術においては、命令キャッシュ、命令バッファ及びブランチターゲットバッファは、互いに別個のチップに作り込まれるか、または1つのチップ上の別個のモジュールとして実装されて来た。不都合なことに、これらの従来技術のやり方では、チップやモジュール同士を相互接続するのに多数のバス回路が必要である。そのために、より大きい面積（回路基板スペースまたはシリコンチップの面積）が必要であり、システムコストが増大する。さらに、より多くのゲートやバッファを通って信号を伝播させる必要があるため、キャッシュアクセスタイムが著しく増大する。

【００１５】本発明によれば、直接マッピング型1次キ

ャッシュに2次キャッシュを統合して作り込んだ効率的なキャッシュメモリシステムが得られる。2次キャッシュを用いると、直接マッピング型キャッシュ本来の効率を保ちつつスラッシングを防止すると共に、ヒット率を改善することができる。本発明によれば、2次キャッシュは好ましくはVLSIチップ上に作られる単一の構造中に直接マッピング型キャッシュと共に集積される。2次キャッシュは、直接マッピング型キャッシュ中に埋込まれ、直接マッピング型キャッシュと同じビット線、センスアンプ及びバスドライバを使用する。また、入力アドレスタグは、1次キャッシュ及び2次キャッシュ中のタグビットと同時に比較される。比較結果が1次キャッシュではミス、2次キャッシュではヒットになると、次のマシンサイクルにおいて2次キャッシュのデータがマイクロプロセッサに供給されるので、主メモリアクセスの必要がない。このように、本発明は、まずデータを直接マッピング型キャッシュへロードするための余分のマシンサイクルを使用することなしに、2次キャッシュからデータを直接読み出すことが可能である。好ましくは、本発明は単一のVLSIチップの形で実施される。一実施例においては、2次キャッシュは、1次キャッシュに対する照会において見つからなかったデータが主メモリからロードされるミスキャッシュを有する。

【００１６】他の実施例においては、2次キャッシュは1次キャッシュの犠牲ラインがロードされる犠牲キャッシュを有する。1次キャッシュの犠牲ラインは、キャッシュミスのデータが主メモリから1次キャッシュにロードされたときに置き換えられるラインである。

【００１７】本発明の他の実施例においては、ブランチ先の命令をロードするのに主メモリにアクセスする必要がなく、分岐命令の効率的な処理が可能な2レベル先取り機構を用いた命令キャッシュが得られる。

【００１８】

【実施例】以下の実施例の詳細な説明においては、説明を明確にするため特定の用語を使用する。しかしながら、本発明は、それらの選択された特定の用語に限定されるものではなく、実質的に同様に作用して実質的に同様の結果を達成する全ての技術的均等物を包含するものである。

【００１９】データキャッシュとして構成した本発明の一実施例のブロック図を図2に示す。この図は、シリコンウェーハまたは半導体チップ上における本発明の構成要素配置について可能な「平面図」の一例を示す。1次キャッシュデータメモリ（1次データキャッシュ、1次キャッシュ）100はシステムの中心部をなしている。好ましくは、1次データキャッシュ100は、2次元に配列された通常のスタティックRAM(SRAM)セルを有する。配列の一方の軸方向には、各々バスインタフェースロジック150に結合された複数のデータビット線112を有し、これらのデータビット線はマイクロプロセッサのデータバス160に結合されている。当技術分野においては周知のように、インタフェースロジック150は、1次キャッシュ100の内容を検出するためのセンスアンプ、1次キャッシュ100のセルに書込み信号を送るための書込みロジック、及びメモリからデータバス160へ供給される出力データを緩衝記憶（バッファー）すると共に増幅するバスドライバ及びデータバッファを有する。これらのインタフェースロジック150の構成要素は当技術分野においては周知である。

【００２０】この1次キャッシュは、直接マッピング型キャッシュとして形成されており、アドレス116をアドレスデコードロジック140に供給することによりアクセスされる。前に図1を参照して説明したように、アドレスデコードロジックは、1次キャッシュ100に下位アドレスビットを供給し、かつインタフェースロジック150の比較器にタグアドレスビット（タグビット）を供給する。さらに、図2の実施例においては、1次キャッシュ100の上またはこれに近接して作り込まれた2次キャッシュ110が設けられている。2次キャッシュは、記憶ライン数が約4～8ラインの小さな完全連想型バッファとして構成することが望ましい。もちろん、2次キャッシュ中のライン数は用途に応じて変えることができる。また、2次キャッシュのメモリセルは、1次キャッシュのセルと同じ構造にすることが好ましい。2次キャッシュのメモリセルには、矢印122により表される複数のマッチ線によって完全連想型比較器120、すなわち比較ロジックが接続されている。比較ロジックは、最初の処理サイクルにおいて、通常の連想メモリ(CAM)セルを用いて入力仮想アドレスまたは物理アドレス124を各CAMセルに記憶されたアドレスと比較する。両者が一致すれば、対応するエントリのマッチ線がアサートされる。次に、直接マッピング型の1次キャッシュ100からキャッシュミスが報告されると、システムは、次のマシンサイクルにおいて2次キャッシュに結合されたミス入力118をアサートする。すると、2次キャッシュのマッチ線が2次キャッシュのワード線（図示省略）をドライブする結果、2次キャッシュからデータが出力される。2次キャッシュからの出力データはビット線112上に出されることによって、直接マッピング型キャッシュでヒットした場合に使用されるのと同じバスインタフェースロジック150を用いてデータバス160へ転送される。

【００２１】図3は、図2の回路の構成要素間の論理的接続関係を示す。図3において、2次キャッシュ110及び1次キャッシュすなわち直接マッピング型キャッシュ100は1つのアドレス入力124に結合されている。アドレスデコーダ（図示省略）によってデコードされた後のアドレスは、3つの論理要素、すなわちタグビット302、ライン番号304及び変位値306を有する。直接マッピング型キャッシュ100にはタグリスト330及び複数のデータライン332が設けられている。アドレス入力のタグビットは線302A

を介して比較器336へ供給される。ライン番号は線304Aを介してタグリストに結合される。当技術分野においては周知のように、この構造の直接マッピング型キャッシュでは、ライン番号を与えてタグリストの中から分離したタグ（分離タグ）334を選択できる。この分離タグは比較器に供給され、分離タグとタグビットが一致すると、比較器はセレクタ338をトリガーして、データライン332の中の1つからデータ項目340を取り出す。データ項目は、変位値306を選択されたラインの始めからのオフセットとして用いることにより選択することができる。次に、データ項目は、データ出力線342を介してマイクロプロセッサの算術論理演算装置(ALU)またはマイクロプロセッサの他の部分に転送される。

【００２２】比較によって一致が得られないと、比較器は線345上にミス信号を発生させる。この信号は、インバータ344に入力して、ハイのミス信号線346をドライブすることができる。また、このミス信号はミスイネーブル回路356をドライブして、マイクロプロセッサに2次キャッシュからデータを読出させる。ミスイネーブル回路は通常の手段によって2次キャッシュに結合されている。

【００２３】2次キャッシュ110へのアクセスは、タグビットを小さなラインの集合の全てのタグと同時に比較するために完全連想型論理を使用する。図3の実施例においては、n組のタグ(タグ1,タグ2、...、タグn)及びライン(ライン1、ライン2、...、ラインn)が示されているが、これらのタグとラインの組数は任意である。このように、2次キャッシュ110は、直接マッピング型キャッシュ100で用いられている2次元配列ではなく、複数のタグ、ライン及び相互接続回路として形成されている。当技術分野においては周知のように、タグとラインの組数を増やすと、処理時間が短くなると共に電力消費が増大する。また、参照番号は、1つのタグ308と1本のライン310にしか示されていないが、他の組についてもこれらと同じ部品が各々タグ及びラインを形成している。

【００２４】2次キャッシュをアドレス指定するために、タグビットは線302Bを介して比較器316へ結合され、同様に他の全てのタグとラインの組に対応する各比較器にも結合されている。当技術分野においては周知のように、完全連想型システムでは、全ての比較器がタグビットを各々の比較器に対応するタグと同時に比較する。すなわち、比較器316は、他の比較器が比較を行なうのと同時に、タグ308をタグビットと比較する。ヒットが起こると（かつミスイネーブル回路356が直接マッピング型キャッシュによるミス後に2次キャッシュをイネーブルしていれば）、比較器はライン310からデータ要素314を抽出するセレクタ318をトリガーする。この場合、いくつかの比較器が同時に比較を実行しているから、図3のいずれかの比較器がヒットを報告し、出力データ要素を供給することができる。データ要素は、変位値をラインの始めからのオフセット312として用いることにより選択することができる。データ要素は、セレクタによりデータ出力線320へ供給され、マイクロプロセッサによって利用される。

【００２５】ミスが発生すると、比較器は、他のミス線と共に3入力NORゲート322に結合されたミス線348をアサートする。全てのミス線がアサートされると、NORゲート322はミス出力線324をアサートし、マイクロプロセッサに2次キャッシュでタグの一致がなかったということを知らせる。

【００２６】直接マッピング型キャッシュからのミス信号及び2次キャッシュからのミス信号は1次ミス出力線354及び2次ミス出力線350を介してANDゲート352に結合される。このように、両方のミス出力線350、354が共にアサートされると、ANDゲートは、大域ミス出力をアサートして、直接マッピング型キャッシュ及び2次キャッシュの両方でミスが発生したので、主メモリにアクセスすることが必要であるということをCPUに指示する。

【００２７】動作について説明すれば、このアーキテクチャによれば、以下のステップによってデータまたは命令を最大2マシンサイクルで直接マッピング型キャッシュまたは2次キャッシュのどちらかから取り出すことが可能である。所与のタグ記憶場所を照会するには、最初のマシンサイクルにおいて直接マッピング型キャッシュのタグリスト330がアクセスされて、直接マッピング型キャッシュにヒットがあるかどうかが確認される。物理タグを有するキャッシュの場合は、タグラインバッファもこのステップでアクセスされる。ヒットが起こると、データは、上記と同じ最初のマシンサイクルにおいて直接マッピング型キャッシュメモリから読み出される。

【００２８】ミスが発生した場合は、2次キャッシュでヒットしていれば次のマシンサイクルにおいて2次キャッシュが読取られる。本発明によれば、2次キャッシュのキャッシュタグを比較するための連想比較動作が、最初のマシンサイクルにおいて直接マッピング型キャッシュにアクセスするために用いられる直接マッピング型比較と同時に行われる。比較器316は比較器336と同時にアクティブ状態を取る。このように、直接マッピング型キャッシュと2次キャッシュが最初のマシンサイクルで同時にチェックされ、直接マッピング型キャッシュでヒットが起こると、その同じ最初のマシンサイクルでデータが読み出される。最初のマシンサイクル中に直接マッピング型キャッシュでミスが起こり、2次キャッシュでヒットが起こると、次のサイクルにおいて2次キャッシュからデータが読み出される。2次キャッシュの読出し動作は、直接マッピング型キャッシュと同じビット線、センスアンプ及びデータバスを使用する。

【００２９】2次キャッシュは犠牲キャッシュまたはミスキャッシュを有することができる。犠牲キャッシュとミスキャッシュの違いは、キャッシュの更新の仕方にあ

る。

【００３０】犠牲キャッシュを用いる場合は、直接マッピング型キャッシュのミスによって置換されるべき直接マッピング型キャッシュ中の犠牲ラインが犠牲キャッシュにコピーされ、このラインは直接マッピング型キャッシュ中では主メモリからの新しいラインにより置換される。

【００３１】ミスキャッシュにおいては、あるミスに対して主メモリから取り出されたラインがミスキャッシュと直接マッピング型キャッシュに同時に記憶される。

【００３２】2次キャッシュで所与のラインがヒットすると、そのラインは2次キャッシュから直接読み出すか、または直接マッピング型キャッシュにロードした後読み出すことができる。最初の場合は、2次キャッシュ中でヒットした各キャッシュアクセスについてデータを読み出すのに2マシンサイクルが必要である。しかしながら、従来技術のシステムと異り、2次キャッシュから直接マッピング型キャッシュへキャッシュラインを転送するための余分のマシンサイクルを必要とすることはない。

【００３３】後の場合においては、例えばカウンタのような寿命追跡または使用追跡機構を用いて2次キャッシュ中のあるラインのアクセス数を監視することができる。そして、この数があるプリセット値を超えたならば、そのラインを将来の直接マッピング型キャッシュのアクセスに備えて、直接マッピング型キャッシュにロードすればよい。この場合、直接マッピング型キャッシュへ転送されたラインは1サイクルだけでアクセスされるが、このラインを転送するのに2サイクルのオーバーヘッドが必要である。そのうちの1サイクルは2次キャッシュからこのラインを読み出すためのサイクルであり、もう1サイクルは逆にそのラインを直接マッピング型キャッシュ書込むためのサイクルである。しかしながら、このラインが2次キャッシュから読み出されたならば、ラインが使用可能になると同時にCPUへ送ることができる。

【００３４】本発明は、命令キャッシュにもデータキャッシュもに使用可能であり、また命令及びデータを共に保持する統合型キャッシュにも使用することができる。また、本発明はミスキャッシュとしても犠牲キャッシュとしても用いることができる。これらのキャッシュの唯一の相違点はキャッシュ中でラインを置換する仕方にある。さらに、本発明はマルチポートキャッシュにも使用することができる。

【００３５】図4及び5には命令キャッシュ（「I-キャッシュ」）の実施例が示されている。この実施例は、1つのVLSI集積回路チップ上に命令バッファ(IB)、ブランチターゲットバッファ(BTB)及び犠牲キャッシュ(VC)を一体に作り込んだ効率的なI-キャッシュである。このハードウェアは、命令がIBまたはBTBからマイクロプロセッサへ出されている間にIBまたはBTBにI-キャッシュから命令を書込むことができる先取り処理が可能である。

【００３６】もう一つの熟慮された実施例においては、図4及び5のI-キャッシュは、第2レベルのキャッシュを有するシステムにおいて第1レベルのキャッシュとして使用することができる。このような実施例においては、図示ハードウェアは、命令がIBまたはBTBから出され続けている間にレベル2からレベル1への先取りが可能である。BTBを使用すると、分岐命令の処理時の遅延を回避することができる。これらを同じI-キャッシュ構造に一体に作り込むと、IB/BTB分離構造の場合と比べて必要表面積が少なくて済み、サイクルタイムが速くなる。この構造は、埋込み型2次キャッシュとして実現することにより、面積の増加を最小限に抑えることができる。

【００３７】図4は、本発明のI-キャッシュとしての実施例の全体的アーキテクチャを示す。この実施例には、別個の2つの命令バス、すなわち読出し命令バス294及び書込み命令バス210（各々「読出しI-バス」及び「書込みI-バス」とも呼ばれる）が設けられている。読出し機能と書込み機能を分離すると、I-キャッシュがレベル2のキャッシュまたは主メモリから書込みI-バスへ命令を先取りするのと同時に、IBまたはBTBから読出しI-バスへ命令を出すことが可能になる。

【００３８】書込みI-バスと読出しI-バスとの間に配置された回路は、図2及び3のデータキャッシュの実施例と構造的にほぼ同じである。入力アドレス224は、第1のアドレスデコーダ240を介して1次命令キャッシュ（1次I-キャッシュ）200に結合され、この1次命令キャッシュはタグラインとデータラインのペアからなる2次元配列として構成することができる。2次I-キャッシュ230は、1次I-キャッシュと結合されると共に、書込みI-バスを介して比較ロジック220と結合されている。比較ロジックは複数のマッチ線225によって2次I-キャッシュ中の記憶セルと接続されている。1次I-キャッシュと2次I-キャッシュは、同じセンスアンプ250、ドライバ290及び分散マルチプレクサ(MUX)292を共用することが望ましい。このように、前述のデータキャッシュ処理と同様にして、これらの構成要素を共用すると、1次I-キャッシュと2次I-キャッシュの間のキャッシュ比較及び命令転送を1マシンサイクルで行うことが可能である。

【００３９】センスアンプ250とバスドライバ290の間には2つの命令バッファ270及び280が構成されている。ブランチターゲットバッファ(BTB)270は、以下に詳細に説明するように、分岐後のターゲット命令のための記憶空間を与える。従って、BTBは、分岐命令が実行された時直ちに将来の命令を実行する準備ができているように、マイクロプロセッサが現命令より論理的に先にある命令を記憶できるようにする。命令バッファ（I-バッファまたはIB）280は、他の将来の命令のためのメモ帳的な空間として用いられる。これらのIB及びBTBは、いずれも

第2のアドレスデコーダ260を介してアドレス224に結合される。これらの各構成要素内部構造は当技術分野においては周知である。

【００４０】本発明のもう一つの実施例においては、比較ロジック及び2次I-キャッシュを使用しない。そのような実施例の動作では、IBが、一杯に書込まれるまでI-キャッシュから間断なく命令を先取りする。これによって、IB中の命令は、マイクロプロセッサが次の命令を要求した時直ちに命令読出しバスへ転送する準備ができる。好ましくは、IB自体は少なくとも2つの命令を記憶し、また先入れ先出し(FIFO)待ち行列として形成される。IBは、待ち行列の先頭から命令を読出しバスへ命令を出す。

【００４１】IBはまだマイクロプロセッサに出されていない将来の命令をチェックする。そして、IBがまだこれから実行しなければならない分岐命令を見つけると、外部のキャッシュコントローラ（図示省略）が命令の先取りを、その分岐に従った命令にアクセスするように切り替える。その後、これらの命令はBTBにロードされる。

【００４２】BTBもFIFO待ち行列の形で実現され、分岐後にターゲット命令を記憶する。分岐命令が実行され、実際に分岐が行われると、その後の命令はBTBから命令読出しバスへ出される。BTBが命令を出している間、BTBが分岐命令を見つけたならば、IBが命令を先取りすることができる。このように、IBとBTBは各分岐毎に互いに役割を切り換えることができる。

【００４３】好ましくは、I-キャッシュからIB及びBTBへフェッチされるワードのサイズは、1命令長より長くする。これによって、いくつかの命令をI-キャッシュからIBまたはBTBへ同時に先取りすることができる。

【００４４】他の実施例においては、外部のキャッシュコントローラに、第2レベルの外部I-キャッシュから1次I-キャッシュへ命令を先取りさせるための手段が設けられる。この第2レベルの外部I-キャッシュは、以下に述べる2次I-キャッシュとは異なる。好ましくは、第2レベルの外部I-キャッシュは、1次I-キャッシュと同じ構造で、適切なカスケードロジックを用いて1次I-キャッシュに結合する。また、好ましくは、1次I-キャッシュからIBまたはBTBへの先取りの間に1次I-キャッシュミスが発生した場合、キャッシュコントローラが第2レベルのI-キャッシュから1次キャッシュへ命令を先取りすることができるようにする。このように、図4の実施例は、第2レベルのI-キャッシュから第1レベルのキャッシュへ先取りを行わせることができる2レベルI-キャッシュ構造になっている。この構造は、ミスを起こした命令が実際に待ち行列の先頭に来るまでIBまたはBTBが命令を出すことができるから、第2レベルのI-キャッシュアクセス中に生じる遅延を隠すのを助ける。

【００４５】キャッシュの下部に図示されている分散マルチプレクサは、複数命令幅IBまたはBTBから一度に1つずつ命令を出す。その結果、キャッシュシステムから出される命令は必ずマイクロプロセッサの命令長と一致する。

【００４６】本発明の一実施例においては、2次キャッシュはI-キャッシュ構造中に埋込まれた犠牲キャッシュまたはミスキャッシュを有する。好ましくは、2次キャッシュは前に図2及び3により説明したように小さな完全連想型バッファを有する。2次キャッシュの比較ロジック220は、2次キャッシュメモリアレイの上に配置することができる。比較ロジックからのマッチ線225は、2次キャッシュメモリアレイのワード線をドライブするよう接続されている。2次キャッシュメモリアレイセル（図示省略）はI-キャッシュセルと同じであることが望ましい。I-キャッシュセルは、好ましくは4トランジスタまたは6トランジスタSRAMセルとする。

【００４７】I-キャッシュの詳細を図5に示す。当業者にとっては明らかなように、図5は、完全なI-キャッシュシステムの並列な全く同じ構造に構成された複数のスライスの中の1枚を示したものである。例えば、図5には、上から下まで、各タイプのデバイスが1つずつしか例示されていない。しかしながら、シリコンチップに実装される完全なI-キャッシュは、図5に例示するスライスと同じスライスを複数並列に有している。

【００４８】I-キャッシュスライスの左半部には、書込み命令バス、書込みドライバ、プリチャージロジック、連想比較ロジック、犠牲キャッシュセル、I-キャッシュセル、及びセンスアンプが形成されている。また、右半部には、命令バッファセル、ブランチターゲットバッファセル、読出しドライバ、分散マルチプレクサ、及び読出し命令バスが形成されている。

【００４９】書込みI-バス210、図5には矢印として示されており、この図に示すスライスから隣の同じスライスへ延びている。書込みI-バス210は、複数の書込みドライバ213を有する書込みドライバ部212に接続されている。書込みドライバ213の出力は、2本の並列ビット線、すなわちビット線214及び否定ビット線216よりなる。当業者であれば、ビット線214と否定ビット線216が互いに反対の論理レベルで動作するということは容易に理解できよう。これらのビット線は、プリチャージロジック部218を介して1次／2次キャッシュアレイに結合される。当技術分野においては周知のように、これらの部分は、書込みI-バス上の命令を1次I-キャッシュ及び／または2次I-キャッシュに記憶させる。

【００５０】比較ロジック220には、2次キャッシュ230の典型的な2次キャッシュ記憶セル232が直接結合されている。この記憶セルは、2次キャッシュワード線234上にデータ出力を送り出す。図示のように、ワード線234は、バスと同様に、どの記憶セルからでも命令ワードを転送することができるよう、隣の記憶セルに接続されている。

【００５１】上記と同じビット線及び否定ビット線には、複数の1次I-キャッシュセル206、208も接続されている。これらのI-キャッシュセル206、208はワード線202、204に各々出力を供給する。図5のＩ－キャッシュの左半部の最下部において、ビット線214、216は、書込みデータ線252に出力データを供給するセンスアンプ(SA)250と結合されている。これらの構成要素の内部構造は当技術分野においては周知である。

【００５２】書込みデータ線252は、図5のI-キャッシュ構造の右半部の最上部にある命令バッファ280と結合されている。図には、各々当技術分野において周知の方法によりFIFO待ち行列の1つの素子として作り込まれた2つの典型的な命令バッファセル280A、280Bが示されている。セル280Aは、読出しワード線（読出しアドレス線）282及び書込みワード線（書込みアドレス線）284を介してアドレス指定され、セル280Bは読出しワード線286及び書込みワード線288を介してアドレス指定される。垂直方向に走る書込みデータ線272及び読出しデータ線274は、セル280A、280Bを相互に接続すると共に、ブランチターゲットバッファ270へ接続する。

【００５３】図5には、2つの典型的なブランチターゲットセル270A、270Bが示されている。これらの各セル270A、270Bは、当技術分野において周知の方法によりFIFO待ち行列の1つの素子として作り込まれる。セル270Aは、読出しワード線276及び書込みワード線278を介してアドレス指定される。セル270Bは、読出しワード線275及び書込みワード線279を介してアドレス指定される。また、これらのセル270A、270Bは、データ線またはビット線272、274によって互いに結合されている。ビット線274の最上端は、当技術分野において周知の方法によりNMOSトランジスタ296を介してプリチャージクロック298に結合されている。この構造によれば、セルの読取りまたは書込みを行う前にビット線をプリチャージすることが可能である。

【００５４】セル270A、270B、280A、280Bから読出された出力データは、読出し出力線291を介して読出しドライバ部290に供給される。これらの出力信号は読出しドライバ部で増幅された後、必ず命令が一度に1つずつ読出し命令バス294に供給されるようにする分散マルチプレクサ(MUX)292へ供給される。

【００５５】本発明は、本願中に具体的に記載した以外の多くの態様で実施することが可能である。従って、本発明の範囲は特許請求の範囲の記載によってのみ決定されるものである。

【００５６】

【発明の効果】本発明によれば、直接マッピング型1次キャッシュに2次キャッシュを統合して作り込んだ効率的なキャッシュメモリシステムが得られ、直接マッピング型キャッシュ本来の効率を保ちつつスラッシングを防止すると共に、ヒット率を改善することができる。

【００５７】また、本発明は、キャッシュミス時にも、まずデータを直接マッピング型1次キャッシュへロードするための余分のマシンサイクルを使用することなしに、2次キャッシュからデータを直接読み出すことが可能であるので、スループットを改善することができる。

【００５８】加えて、本発明をデータキャッシュとして使用した場合には、直接マッピング型キャッシュを埋込み式の犠牲／ミスキャッシュと共に用いることによりスラッシングが防止されること、サイクルタイムが改善されること、電力消費が小さいこと、面積効率がよいこと、及びセンスアンプのスイッチングノイズに対する感度が低くなることなどの長所がある。更に、本発明を命令キャッシュとして使用した場合には、上記全ての長所に加えて、2レベルI-キャッシュ先取り機構が得られること、及びI-キャッシュを命令バッファとブランチターゲットバッファに結合するための先取り機構が得られることなどの長所がある。これらの長所によって、キャッシュアクセスのサイクルタイムが短縮され、ヒット率がより高くなり、またシリコンチップの面積を最小限にすることが可能となる。

【図面の簡単な説明】

【図１】従来技術として知られている直接マッピング型キャッシュメモリのブロック図。

【図２】本発明の直接マッピング型1次キャッシュと完全連想型2次キャッシュメモリシステムのブロック図。

【図３】図2の1次及び2次キャッシュについてのより詳細な構造図。

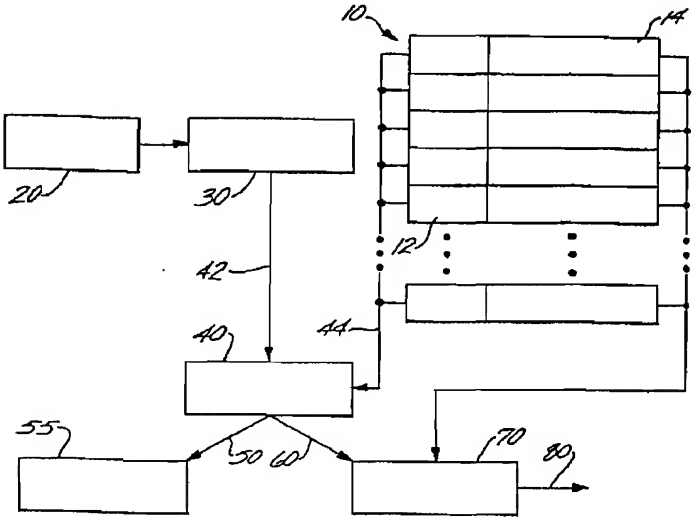
【図４】本発明のもう一つの実施例である命令キャッシュメモリシステムのブロック図。

【図５】図4の命令キャッシュにおける複数のスライスの一つについてのより詳細な構造図。

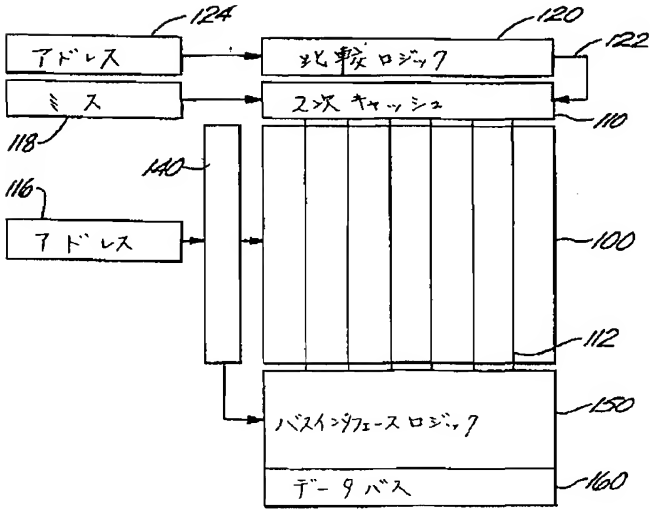
【符号の説明】

１００：直接マッピング型1次キャッシュ
１１０：2次キャッシュ
１１６：アドレス
１１８：ミス入力
１２０：完全連想型比較器（比較ロジック）
１２２：マッチ線
１２４：アドレス入力
１４０：アドレスデコードロジック
１５０：バスインターフェースロジック
１６０：データバス

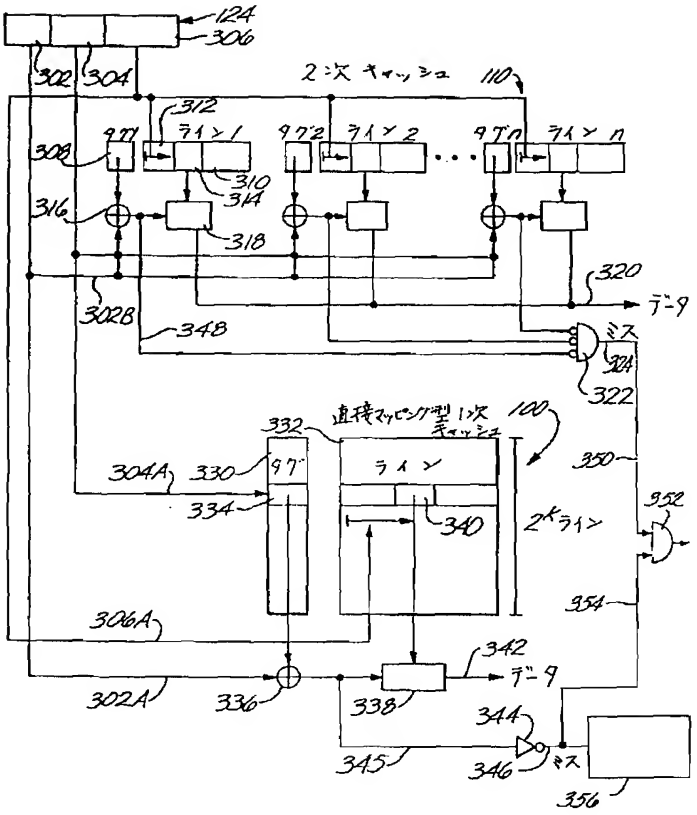
【図１】


10
14
20
30
12
42
44
40
55
50
60
70
80

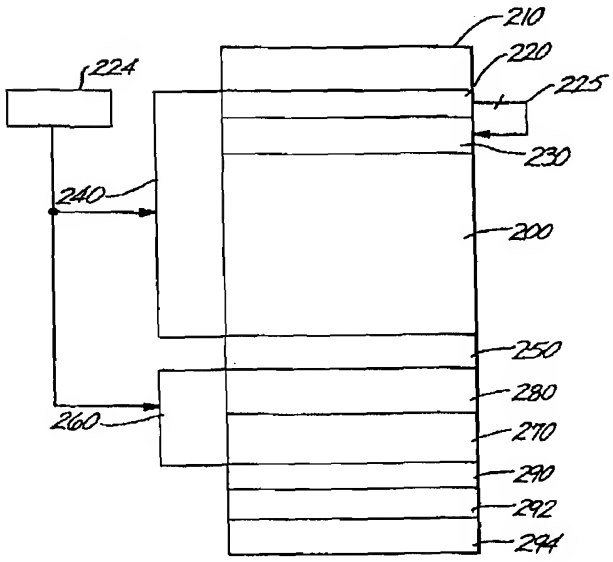
【図２】


124
120
122
アドレス
比較ロジック
ミス
2次キャッシュ
110
118
140
116
アドレス
100
112
バスインタフェースロジック
150
データバス
160

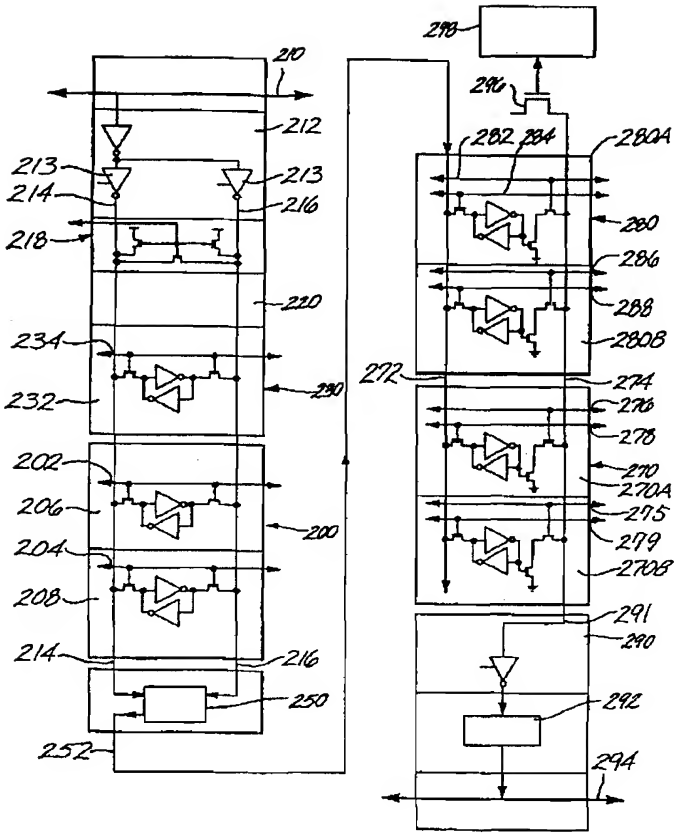
【図3】


124
306
302
304
2次キャッシュ
110
312
タグ1
ライン1
タグ2
ライン2
タグn
ライン n
308
310
314
316
318
320
データ
302B
348
ミス
324
322
直接マッピング型1次キャッシュ
100
332
304A
330
タグ
ライン
334
340
2^k ライン
350
352
354
306A
342
データ
302A
336
338
344
345
346
ミス
356

【図4】


210
220
224
225
230
240
200
250
280
260
270
290
292
294

【図５】


298
210
296
212
282
284
280A
213
213
214
216
280
218
286
220
288
280B
234
272
274
290
276
232
278
202
270
270A
206
275
200
279
204
270B
208
291
214
216
290
250
292
252
294